We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter 9

Conclusion: Towards Achievable and Sustainable Open Scientific Data

Vera J. Lipton

This chapter summarises the findings of the study by answering the research questions posed. The chapter consists of four parts:

1. Vision: What are the expected benefits associated with the curation and release of open scientific data?
2. Policy: What is the scope of the open data policies recently introduced by research funders and publishers?
3. Practice: How are selected data-centric public research organisations implementing open data? What are the legal and other challenges emerging in the process of implementation? Is open scientific data an achievable objective?
4. A way forward: What can be done to promote open access to scientific data across different research disciplines? Is there a need to revise the open data mandates?

Introduction

This book began with the call from research funders and publishers for increased access to research data so as to facilitate its increased uptake and reuse by others. The principal triggers for the renewed emphasis on sharing research data are the open data policies introduced by research funders and publishers in many jurisdictions in the world.

In this final chapter, we take a step back to review the findings of this study to evaluate the effect of these policies on the practice of data sharing as open data. I start with an overview of the expected benefits of open scientific data and the assumptions that led governments to introduce the policies. This is followed by a summary of the scope of the mandates and then an outline of the challenges associated with the practice of open data at CERN and in clinical trials. The final section briefly summarises the staged model for open scientific data introduced in the previous chapter.

9.1 Vision: what are the expected benefits associated with the curation and open release of scientific data?

The research data landscape has changed considerably in recent years. The open data policies introduced by research funders and publishers since 2010 have created

a momentum, driving research data curation and release globally, this book finds.[1] Open data is developing concurrently with the open publications sector, which has accelerated the speed and ease of making research publications freely available in digital formats.[2] The last few years have also seen the emergence of data journals and discipline-specific data repositories that enable researchers to deposit their research data along with publications.[3]

These developments are underpinned by the strong endorsements of open data practice by major public research funders—including the National Institutes of Health in the United States, the European Commission,[4] stringent regulatory authorities such as the European Medicines Agency (EMA),[5] and esteemed research organisations such as CERN and NASA, among many others. These actors have championed open scientific data and are developing major infrastructures for data deposit and discoverability.

Implicit in these developments is the understanding of the common objectives and benefits of open scientific data—to advance and democratise science by increasing the uptake and reuse of scientific knowledge and data, to increase the quality and transparency of published scientific results, to enable the verification and reproducibility of scientific results, and to facilitate the continuing shift towards digital modes of science production and dissemination.[6] Also implicit in these benefits is the desire to find solutions to some of the biggest challenges facing humanity and the planet today—global warming, food security and poverty, the insatiable demand for energy and resources, increased pollution, growing urbanisation, and the quest for increased knowledge, longevity, and an improved quality of life that increasingly depend on the application of science and technology.[7]

In this world of rapid technological changes, in which scientific knowledge increasingly means power and market advantage,[8] the demand for scientific knowledge and data is also increasing.[9] While most research remains publicly funded,[10] recent years have seen an uptake of open innovation strategies by companies—especially those that source knowledge from external sources,[11] as evidenced in growing demand for collaborations and partnerships with universities and public research organisations.[12] Such partnerships and innovation strategies have resulted in the increased commoditisation of science by businesses—a trend that is especially evident in the biological and medical sciences as well as in engineering.[13]

In this context, scientific data in the public domain has the potential to impact the economic context in which power and control over science are distributed in society. Open scientific data ensures that the outcomes of public science remain available without any restrictions and for reuse by anyone, including future generations of researchers working in the public and private sectors, and anywhere in the

---

[1] See conclusion in Chapter 3 of this book.

[2] See Chapter 3, Sections 3.2 and 3.3, and Chapter 8, Sections 8.2 and 8.3.1.

[3] See Chapter 3, Section 3.2.

[4] *Ibid.*

[5] See Chapter 6, Section 6.2 and Chapter 7, Section 7.5.3.

[6] See Chapter 2, Sections 2.3 and 2.4.

[7] See Chapter 1, Section 1.1.1.

[8] See Chapter 2, Section 2.2 and Footnote 57.

[9] *Ibid*, Footnote 57.

[10] See Chapter 2, Section 2.5, Footnote 99.

[11] See Chapter 2, Section 2.5.

[12] *Ibid*, Footnote 100.

[13] See Chapter 2, Section 2.2 and Chapter 1, Section 1.1.1, Footnote 33.

world. Such a practice brings about huge economic benefits for countries that invest in the development of open data, as evidenced in the Human Genome Project and the Global Positioning System (GPS)—two early, large-scale open data initiatives.

The Human Genome Project cost the US government US$3.8 billion to develop and up until today has generated around US$750 billion in biotechnology industry output in that country.[14] Compare US$750 billion with just over US$1 billion received from biotechnology licencing revenue by the top 15 universities in the United States and just over US$400 million of commercial income received from IP licencing by that country's biomedical research institutes.[15]

The economic benefit that the United States has accrued from GPS technology up until 2013 was estimated at about US$56 billion.[16] Compare this with the less than US$3 billion received as income from the commercialisation of research by all universities in the United States in 2016,[17] with over 85% of universities finding themselves unable to realise enough income to cover the costs of running their technology transfer offices.[18]

The economic justification of innovation is clearly on the side of open data, and governments should not be afraid to invest in the development of open technologies. The potential benefits for local economies are enormous.

9.2 Policy: what is the scope of the open data policies recently introduced by research funders and publishers?

Research funders and publishers have played a critical role in driving open scientific data. Beyond federal governments, private not-for-profit research funders such as the Bill and Melinda Gates Foundation and Wellcome Trust have adopted open data policies. These policies have changed the game and have, within a span of around 5 years, led researchers and their organisations to curate, document, and share their data.[19] Most funders have some form of policy regarding *research data management* (RDM)—ranging from requiring data management plans at the proposal stage through to expectations about depositing and sharing data. In response to these policies, research organisations have developed or strengthened internal RDM functions.

These policy adjustments vest the responsibility for data curation and release in researchers. The policies vary in their scope and in the specific requirements for sharing research data. Some policies 'recommend' or 'strongly encourage' data sharing, while others 'require' it. Several policies explicitly 'mandate' data sharing for research that receives grant money and stipulate requirements on when, how, and what data should be deposited and where.[20] The Public Library of Science mandates data availability as a condition of publication. Other journals, such as *Nature* and *Science*, expect researchers who publish within their pages to provide data 'on request', without requiring the deposit of data on the date of publication.

The first evaluations of these policies have found a strong correlation exists between the existence of data policies and data deposit practice.[21] Another

---

[14] Chapter 2, Section 2.5, under Human Genome Project.

[15] See Chapter 2, Section 2.5, Footnote 124.

[16] *Ibid*, under Global Positioning System.

[17] See Chapter 2, Section 2.5, Footnote 124.

[18] *Ibid.*

[19] See Chapter 3, Conclusion, and Sections 3.2 and 3.3.

[20] See Chapter 3, Section 3.4.

[21] See Chapter 6, Section 6.1, Footnote 15.

important finding is that more prescriptive policies—those with a mandate for data deposit along with a statement on data sharing included in the manuscript, have achieved the greatest deposit rates.[22]

However, a major theme that has emerged in this book is that the meaning of 'research data' varies across scientific disciplines, across various levels of data processing, and can originate from many different sources.[23] In addition, research practices vary widely across scientific disciplines and so does the collection and preparation of open access data.

The inability to clearly acknowledge and articulate the heterogenous nature of research data is a major shortcoming of the open data mandates; this book has argued.[24] In particular, the opening up of research data requires adopting an open mindset and the acknowledgement that 'one size does not fit all'; a mindset that finds RDM is an ongoing process that is as important a driver of improved science as is the resulting open data. Another key finding is that the quality of open data is far more important than quantity. More open scientific data, by itself, does not necessarily lead to more open science, more easily reproducible science, or improved and data-driven science.

This study cautions against any standardised approach to defining 'data'. While such approaches have generally proved useful when developing open access to publications, such approaches are neither suitable nor appropriate for open scientific data, this book argues.[25] If open scientific data is to be sustainable, then cultural, research practice, and organisational issues must first be addressed.

Yet librarians and research funders, who play pivotal roles in facilitating open access to scientific publications, tend to apply the same 'standardised' principles and approaches to research data. In particular, many librarians are calling for the standardisation of research data formats and metadata descriptors for inclusion in the policies of research funders and publishers.[26] This creates confusion and challenges for researchers, who are required to comply with the mandates introduced by research funders but are unable to do so because the complexity and heterogenous nature of open data simply makes it impossible for them to apply the same set of rules to every research project and dataset.

Common language and search structures can indeed facilitate discoverability of data. However, every dataset is unique, requiring different languages to describe the data and provide all supporting documentation, software, algorithms, and metadata so as to facilitate the reuse of the data. In this sense, research data is more analogous to archival materials rather than to open publications.

The experience from CERN is that only researchers can develop the necessary data descriptors and that these descriptors need to be rigorously tested and embedded in research practice before any common language and data structures can be contemplated.[27] In other words, attempts at research data standardisation need to be driven bottom-up, by researchers. External approaches that would impose common descriptors on research data would be unhelpful unless the descriptors are already firmly established in research practice. Given the recent and novel nature of open scientific data, such pilots are only just now starting to emerge. The notion of research data and its structuring and sharing require more refinement.

---

[22] *Ibid*, Footnote 17.

[23] See Chapter 4, Sections 4.1 and 4.2.

[24] See Chapter 4, Conclusion.

[25] See Chapter 8, Sections 8.1 and 8.2

[26] See Chapter 3, Section 3.4.

[27] See Chapter 5, Sections 5.3 and especially 5.3.4.

DOI: http://dx.doi.org/10.5772/intechopen.87631

In the meantime, open data as a default practice seems appropriate for data underpinning scientific publications—to facilitate the validation of results. Yet 'open by default' is not, at this stage, feasible for data produced in clinical trials and data collected in particle physics experiments, even though well-documented and well-curated digital data, including raw data and metadata, is generally available. Most of the data can only be shared with expert collaborators. Carefully selected subsets of the data are, however, increasingly becoming available as open data for educational purposes. Open data is also paving the way for making scientific experiments more accessible to wider audiences.

9.3 Practice

9.3.1 How are selected data-centric public research organisations implementing open data? Is open scientific data an achievable objective?

In assessing early experiences with open scientific data at CERN and with clinical trial data, this book finds that curating scientific data for public release is far more complex and costly than governments and research funders had envisaged.

The major complication is that implementing open scientific data requires appropriate RDM. Public research organisations in general, and universities in particular, have very limited experience in this area.[28] Furthermore, the key stakeholders in the process have different, often conflicting, interests, and concerns about research data.

For researchers, the need to ensure the ethics and validity of secondary data analyses and the recognition of their efforts vested in data curation are the most prominent concerns. From the perspective of research sponsors and publishers, safeguarding their economic interests through intellectual property and confidentiality remain important considerations that directly challenge the practice of open scientific data.

Understanding the requirements for responsible data sharing and ensuring compliance with these requirements pose fresh challenges to research organisations. Maintaining the privacy of subjects involved in data collection, particularly in clinical trials, is an additional concern for medical research institutes. Furthermore, digital curation of research data is labour and resource-intensive and requires substantial investments in data infrastructures and new business models. In this context, many research organisations point out that open scientific data should not be an unfunded mandate. This is particularly the concern among researchers collecting clinical trial data, who fear that the funding needed for data curation will diminish the resources available to conduct new trials.[29]

The lessons learnt with implementing open data at CERN can prove helpful to other research organisations active in different areas of science. One particular area of emerging best practice is that the implementation of open data within organisations needs to be embraced and discussed by all—researchers, management and librarians. At CERN, such discussions were initiated through the development of internal open data preservation and sharing policies within the four main research teams. The vigorous debate that occurred at many different levels during the process transformed the whole organisation, including its conduct of data-driven research.[30] The resulting open data policies have created a shared understanding of

[28] See Chapter 5, Section 5.1.

[29] See Chapter 6, Section 6.2.4, Footnotes 72 and 73.

[30] See Chapter 5, Sections 5.3, and Conclusion.

the processes leading to data reusability and established the potential for data sharing with external users.[31]

One particular finding at CERN was that the value of open scientific data lies primarily in its quality, determined by two factors—robust data management practices within organisations and the potential in open data for future use and reuse.[32] From this came the development of the Open Data portal at CERN.

Despite these learnings and insights, CERN has not yet made available as open data all the data it produces. It has divided prospective users into four groups—ranging from a base level, offering direct access by anyone to the data underlying publications, through to the restricted access to the entire raw dataset only available to selected expert collaborators. This user hierarchy is necessary because CERN does not, at this time, have the data-processing capacity to accommodate universal and unrestricted access and also because some of the data requires knowledge of particle physics to understand and reuse it.[33]

In medicine, the sharing of clinical trial and genomic data has been an established practice for several years. It has gained new momentum with the release of open data mandates by research funders, by publishers, and especially by the EMA. New requirements for data sharing have also lead to greater transparency and increased data sharing in industry. Open sharing of data submitted to regulatory authorities has been tested by courts, which have upheld, in all cases, the open approach championed by the EMA.[34]

The key consideration in sharing patient-level data as open data is the protection of privacy and confidentiality. Research organisations have dealt with these concerns for many years and have in place well-tested procedures for research ethics along with data sharing protocols.[35] These are supported by the rigorous training of researchers, including the certification of researchers who collect and work with data involving human subjects.

However, recent unauthorised data sharing and privacy breaches by several large companies have brought renewed attention from policymakers to ensuring data privacy and confidentiality. As the result of the widespread publicity for privacy breaches at companies such as Facebook and Yahoo, policymakers are seeking to interfere with established decentralised research practice and to institute centrally controlled mechanisms to manage the privacy and confidentiality of data, with the vetting of prospective users.[36] In particular, this is the policy approach adopted by governments in the United Kingdom, New Zealand and, very recently, in Australia. There is a proposal to apply this approach to research data and, on first sight, it appears that it would apply to all research data.[37]

Such centralised approaches are unlikely to yield the desired economic and social benefits that open data presents. If there is one lesson learnt from the remarkable growth of the biomedical industry in Europe and the United States, it is that decentralised and open research can accelerate the pace of discovery and innovation, fuel economic growth and strengthen global competitiveness. This potential can only be realised if research data is available broadly and is reused by others.

---

[31] *Ibid.*

[32] See Chapter 5, Conclusion.

[33] See Chapter 5, Section 5.2.2.

[34] See Chapter 7, Section 7.5.3.

[35] See Chapter 6, Sections 6.2.1 and 6.2.2, and Chapter 7, Section 7.5.

[36] See Chapter 7, Section 7.5.4.

[37] *Ibid.*

Another important issue that has emerged in the implementation of open data, both at CERN and in clinical trials, is the necessity to define the levels of processing and other parameters that can make data reusable by others. Best practice in both fields confirms that research data, software and metadata—the three components of research data generally specified in the policies of research funders and publishers—are not sufficient to enable independent data reuse.[38]

Also required is a detailed description of the assumptions made by the original data collectors during the different stages of their research and data analysis, along with the statistical methods used to clean, process, aggregate and analyse the data. Such steps are rarely recorded as part of research practice and more study is needed to determine the scope and level of documentation required to achieve data reusability across different scientific disciplines and projects.

With this in mind, it is important for research funders across the different scientific disciplines to ascertain the levels at which scientific data is generally collected and processed across each scientific discipline. Funders should then set the boundaries for the levels at which the data holds the highest potential for reuse by others, whether as researchers (expert users) or as other interested users.

In addition, there is the need for reconsideration of the calls by research funders and policymakers for research reproducibility. This study finds that sharing of research data as open data does not necessarily or easily lead to research reproducibility. Low-level data (raw data) are generally required for this purpose and such data may not be readily available for sharing as open data or they can be costly to curate. Even where low-level data and all supporting analyses, algorithms and software are meticulously documented and are made available, experts in the same field of science may not achieve duplicate results by reusing the same data and applying the same techniques.[39]

Moreover, reproducibility studies can be costly, as evidenced in clinical trials and experienced first-hand by biomedical companies trying to replicate the research of competitors.[40] Therefore, reproducibility should only be the desired and stated objective in carefully selected research areas or research projects—such as those designed by drug regulators or those commissioned by courts to verify ambiguous claims made by pharmaceutical companies in their marketing applications for the approval of new products. Reproducibility should not be held as the 'golden standard' for science,[41] and it should not be one of the key objectives for open scientific data that research funders advocate.

9.3.2 What are the legal and other challenges emerging in the process of implementation?

Depositing research data in the public domain has highlighted the need to determine the legal owner of the dataset.[42] Uncertainties around the application of copyright to the various forms of data, and around data ownership in the research sector and in academia, have been identified as the root causes of subsequent problems affecting data licencing and the lack of clarity around conditions governing data reuse.[43]

---

[38] See Chapter 8, Section 8.3.6.

[39] See Chapter 6, Section 6.4.3.

[40] *Ibid*.

[41] *Ibid*.

[42] See Chapter 7, Section 7.2 and Conclusion and Chapter 8, Section 8.3.8.

[43] *Ibid.*

A further concern is the duty of fidelity that researchers have to their employers, which may prevent them from disclosing information acquired in the course of their employment.[44] The duty of confidentiality may also arise in collaborations with private sector sponsors. This book recommends more analysis of the relationship between the ownership of research data, in its different forms, and the interplay of that with possible copyright protection and confidentiality issues.[45]

Reuse of open data can give rise to legal problems, especially in the context of text and data mining, which is necessary to extract value and insight from datasets.[46] Since data mining typically requires the making of a (temporary) copy of the dataset, it is likely that the act of copying would amount to copyright infringement.

In this matter, compared with their counterparts in the United States, Europe and in other parts of the world, Australian research organisations seem disadvantaged. Such an inhibition for text and data mining also makes Australia a less attractive destination for data-driven businesses. This study proposes introducing a text and data mining exemption into the Australian *Copyright Act 1968* [496].[47]

9.3.3 Is open scientific data an achievable objective?

Taken together, the lessons learnt from the implementation of open scientific data—along with the financial and research benefits accrued from open data to this point, the potential future benefits and the increased need in the digital era for researchers to gain faster access to research data to conduct research—lead to the conclusion in this book that open scientific data is indeed an achievable objective and should become a priority for all research organisations.

However, curating all publicly funded research data as open data is not possible with current technology, nor it is currently achievable at recoverable cost. There remain necessary choices about what data to select for curation and release as open data.

The staged model proposed in this book offers some suggestions on how choices can be made so as to balance the individual responsibilities of researchers for curating research data with the collective benefits likely to accrue to other researchers and to society through reuse of the data.

9.4 A way forward: what can be done to promote open access to scientific data across different research disciplines? Is there a need to revise the open data mandates?

This book found that the open data mandates as they stand today do not acknowledge the diversity of research data as it occurs across different research disciplines and at different stages of processing and control. No uniform answers exist for the question of what defines data, and therefore, it is also difficult to determine what data is worth preserving into the future.

The staged model proposed in the preceding chapter encourages research organisations to define the content of the data they hold as well as to define the stages of its processing. This model, along with eight recommendations, presents a roadmap towards more achievable and sustainable open scientific data.

---

[44] See Chapter 7, Section 7.5 and Chapter 8, Section 8.3.8.

[45] *Ibid.*

[46] See Chapter 7, Section 7.4 and Chapter 8, Section 8.3.9.

[47] *Ibid.*

DOI: http://dx.doi.org/10.5772/intechopen.87631

The proposed model includes four levels of data processing and release. It calls for default and immediate open access to data that underpins results published in scientific publications (Level 1 data). The staged model recognises the value open that data can deliver if it is used for educational and outreach purposes, as demonstrated with Level 2 data at CERN. The proposed model also recognises that not all research data can be of interest to the general public and that there are certain risks associated with sharing some types of data. Therefore, the model proposes that Levels 3 and 4 data may not be shared immediately after the publication of research results and that such data should be restricted for reuse by expert users with relevant competence.[48]

The factors that drive the independent reuse of open data are not known at this stage and will emerge over time as open data collections increase and gain in value. For now, open data practice may not be easy to implement, yet the individual and organisational lessons learnt are significant discoveries on the transformational journey to digital science.

As technologies evolve and as our ability to work with open data increases, the value of open data will increase also. Those governments, researchers, and organisations that learn to share their research data, and that learn to harness the value of data released by others, will become the visionaries to lead us into a data-enriched future.

> *"We must believe that we are gifted for something, and that this thing, at whatever cost, must be attained."*
>
> *Marie Skłodowska-Curie*

[48] See Chapter 8, Section 8.3.

Author details

Vera J. Lipton
Zvi Meitar Institute for Legal Implications of Emerging Technologies,
Harry Radzyner Law School, IDC Herzliya, Israel

*Address all correspondence to: vera.lipton@bigpond.com